Vida: How to use Bayesian inference to de-anonymize persistent communications

George Danezis[1] and Carmela Troncoso[2]

[1] Microsoft Research Cambridge
gdane@microsoft.com
[2] K.U. Leuven/IBBT, ESAT/SCD-COSIC
Carmela.Troncoso@esat.kuleuven.be

Abstract. We present the *Vida* family of abstractions of anonymous communication systems, model them probabilistically and apply Bayesian inference to extract patterns of communications and user profiles. The first is a very generic Vida Black-box model that can be used to analyse information about all users in a system simultaneously, while the second is a simpler Vida Red-Blue model, that is very efficient when used to gain information about particular target senders and receivers. We evaluate the Red-Blue model to find that it is competitive with other established long-term traffic analysis attacks, while additionally providing reliable error estimates, and being more flexible and expressive.

1 Introduction

Anonymous communications allow conversing parties on a network to exchange messages without revealing their network identifiers to each other or to third party observers. Anonymity is of special importance to ensure privacy, support protocols such as on-line polls, or enable high-security government or military communications over commodity network infrastructures.

The most practical proposal for engineering anonymous communications is the *mix*, proposed by David Chaum [3] in 1981. A mix is a network router offering a special security property: it hides the correspondences between its input and output messages, thus providing some degree of anonymity. A large body of research, surveyed in [6], is concerned with extending and refining mix based protocols.

In parallel with advances in anonymity, techniques have been developed to uncover persistent and repeated patterns of communication through mix networks. Such attacks were first named "intersection attacks" [17] since they were based on the idea that when a target user systematically communicates with a single friend it is possible to uncover the identity of the latter by intersecting the anonymity sets of the sent messages. Kesdogan *et al.* [1, 12, 13] introduced a family of disclosure and hitting set attacks that generalises this idea to users with multiple friends. These attacks' result is the set of friends of each sender being uncovered, after a number of messages communicated. Statistical variants of these attacks were also developed, known as statistical disclosure attacks [5],

and applied to pool mixes [8], traffic containing replies [7], and evaluated against complex models [15]. The state of the art in statistical disclosure is the Perfect Matching Disclosure Attack introduced by Troncoso *et al.* [21]. The PMDA allows to guess who are communication partners in a round of mixing with higher accuracy than its predecessors. Further the authors show how this information can be in turn used to improve the estimation of users' sending profiles.

This work re-examines the problem of extracting profiles and, in parallel, uncover who is talking with whom, from traffic traces of anonymous communications. We offer a generalisation of the disclosure attack model of an anonymity system [1, 12, 13], and analyse it using modern Bayesian statistics. We note that at the heart of long term traffic analysis lies an inference problem: from a set of public observations the adversary tries to infer a "hidden state relating" to who is talking to whom, as well as their long term contacts. Applying Bayesian techniques provides a sound framework on which to build attacks, standard well studied algorithms to co-estimate multiple quantities, as well as accurate estimates of error.

Our key contributions are first *the very generic* Vida *models to represent long term attacks against any anonymity system*, and second *the application of Bayesian inference techniques to traffic analysis*. Throughout this work we show that our models and techniques lead to effective de-anonymization algorithms, and produce accurate error estimates. Furthermore they are far more flexible and reliable than previous ad-hoc techniques.

This paper is organised as follows: Sect. 2 offers an overview of Bayesian inference techniques, their relevance to traffic analysis, as well as an overview of the Gibbs sampling algorithm; Sect. 3 presents the Vida generic model for anonymous communications, that can be used to model any system. In Sect. 4 we present a simplification of the model, the Vida Red-Blue model, that allows an adversary to perform inference on selected targets, as it would be operationally the case, along with an evaluation of the effectiveness of the inference technique. Finally we discuss the future directions of inference and traffic analysis in Sect. 5 and conclude in Sect. 6.

2 Bayesian inference and Monte Carlo methods

Bayesian inference is a branch of statistics with applications to machine learning and estimation [14]. Its key methodology consists of constructing a full probabilistic model of all variables in a system under study. Given observations of some of the variables, the model can be used to extract the probability distributions over the remaining, hidden, variables.

To be more formal lets assume that an abstract system consists of a set of hidden state variables $\mathcal{HS}$ and observations $\mathcal{O}$. We assign to each possible set of these variables a joint probability $\Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}]$ given a particular model $\mathcal{C}$. By applying Bayes rule we can find the distribution of the hidden state given the

observations as:

$$\Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}] = \Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}] \cdot \Pr[\mathcal{O}|\mathcal{C}] \Rightarrow \Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}] = \frac{\Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}]}{\Pr[\mathcal{O}|\mathcal{C}]} \Rightarrow$$
$$\Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}] = \frac{\Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}]}{\sum_{\forall \mathcal{HS}} \Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}] \equiv \mathcal{Z}} = \frac{\Pr[\mathcal{O}|\mathcal{HS}, \mathcal{C}] \cdot \Pr[\mathcal{HS}|\mathcal{C}]}{\mathcal{Z}}$$

The joint probability $\Pr[\mathcal{HS}, \mathcal{O}|\mathcal{C}]$ is decomposed into the equivalent $\Pr[\mathcal{O}|\mathcal{HS}, \mathcal{C}] \cdot \Pr[\mathcal{HS}|\mathcal{C}]$, describing the model and the a-prior distribution over the hidden state. The quantity $\mathcal{Z}$ is simply a normalising factor.

There are key advantages in using a Bayesian approach to inference that make it very suitable for traffic analysis applications:

- The problem of traffic analysis is reduced to building a generative model of the system under analysis. Knowing how the system functions is sufficient to encode and perform the attacks, and the inference steps are, in theory, easily derived from this forward model. In practice computational limitations require careful crafting of the models and the inference techniques to be able to handle large systems.
- The Bayesian approach allows to infer as many characteristics of the system as needed by introducing them in the probabilistic model. This permits to infer several hidden variables jointly as we show for users' sending profiles and their recipient choices for each message.
- A Bayesian treatment results in probability distributions over all possible hidden states, not only the most probable one as many current traffic analysis methods do. The marginal distributions over different aspects of the hidden state can be used to measure the certainty of the attacker, and provide good estimates of her probability of error.

The last point is the most important one: the probability distribution over hidden states given an observation, $\Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}]$, contains a lot of information about all possible states. When traffic analysis is used operationally the probability of error of particular aspects of the hidden state can be calculated to inform decision making. It is very different to assert that, in both cases, the most likely correspondent of Alice is Bob, with certainty 99% versus with certainty 5%. Extracting probability distributions over the hidden state allows us to compute such error estimates directly, without the need for an ad-hoc analysis of false positives and false negatives. Furthermore, the analyst can use the inferred probability distribution to calculate directly anonymity metrics [9, 19].

Despite their power Bayesian techniques come at a considerable computational cost. It is often not possible to compute or characterise directly the distribution $\Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}]$ due to its complexities. In those cases sampling based methods are available to extract some of its characteristics. The key idea is that a set of samples $\mathcal{HS}_0, \ldots, \mathcal{HS}_\iota \sim \Pr[\mathcal{HS}|\mathcal{O}, \mathcal{C}]$ are drawn from the a-posterior distribution, and used to estimate particular marginal probability distributions of interest. For this purpose, Markov Chain Monte Carlo methods have been proposed. These are stochastic techniques that perform a long random walk

on a state space representing the hidden information, using specially crafted transition probabilities that make the walk converge to the target stationary distribution, namely $\Pr[\mathcal{HS}|\mathcal{O},\mathcal{C}]$. Once the Markov Chain has been built, samples of the hidden states of the system can be obtained by taking the current state of the simulation after a certain number of iterations.

2.1 Gibbs sampler

The Gibbs sampler [11] is a Markov Chain Monte Carlo method to sample from joint distributions that have easy to sample marginal distributions. These joint distributions are often the a-posterior distribution resulting from the application of Bayes theorem, and thus Gibbs sampling has been extensively used to solve Bayesian inference problems. The operation of the Gibbs sampler is often referred to as *simulation*, but we must stress that it is unrelated to simulating the operation of the system under attack.

For illustration purposes we assume an a-posterior distribution $\Pr[\mathcal{HS}|\mathcal{O},\mathcal{C}]$ can be written as a joint probability distribution $\Pr[X,Y|\mathcal{O},\mathcal{C}]$ that is difficult to sample directly. If, on the other hand, there is an efficient way of sampling from the marginal distributions $\Pr[X|Y,\mathcal{O},\mathcal{C}]$ and $\Pr[Y|X,\mathcal{O},\mathcal{C}]$, then Gibbs sampling is an iterative technique to draw samples from the joint distribution $\Pr[X,Y|\mathcal{O},\mathcal{C}]$. The algorithm starts at an arbitrary state (x_0,y_0). Then it iteratively updates each of the components through sampling from their respective distributions, i.e. $x_i \sim \Pr[X|Y=y_{i-1},\mathcal{O},\mathcal{C}]$, and $y_i \sim \Pr[Y|X=x_i,\mathcal{O},\mathcal{C}]$. After a sufficient number of iterations, the sample (x_i,y_i) is distributed according to the target distribution, and the procedure can be repeated to draw more samples. We note that in this process the computation of the normalising factor $\mathcal{Z}$ is not needed.

The other parameters of the Gibbs algorithm, namely the number of iterations necessary per sample, as well as the number of samples are also of some importance. The number of iterations has to be high enough to ensure the output samples are statistically independent. Calculating it exactly is difficult so we use conservative estimates to ensure we get good samples. The number of samples to be extracted, on the other hand, depends on the necessary accuracy when estimating the marginal distributions, which can be increased by running the sampler longer.

3 The Vida general Black-box model for anonymity systems

Long term attacks traditionally abstract the internal functioning of any anonymity system and represent it as an opaque router, effectively operating as a very large threshold mix. This model has its limitations, and some studies have attempted to extend it. In this section we first propose the Vida Black-box model, the most flexible abstraction of an anonymity system so far, and base our Bayesian analysis on this model.

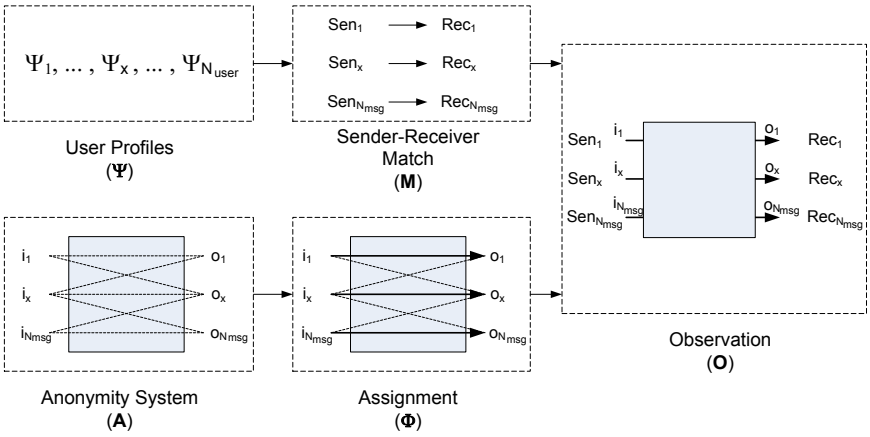

Fig. 1. The generative model used for Bayesian inference in anonymous communications.

We start by proposing a 'forward' generative model describing how messages are generated and sent through the anonymity system. We then use Bayes rule to 'invert' the problem and perform inference on the unknown quantities. The broad outline of the generative model is depicted in Figure 1.

An anonymity system is abstracted as containing N_{user} users that send N_{msg} messages to each other. Each user is associated with a sending profile Ψ_x describing how they select their correspondents when sending a message. We assume, in this work, that those profiles are simple multinomial distributions, that are sampled independently when a message is to be sent to determine the receiver. We denote the collection of all sending profiles by $\Psi = \{\Psi_x | x = 1 \dots N_{\text{user}}\}$.

A given sequence of N_{msg} senders out of the N_{user} users of the system, denoted by $\text{Sen}_1, \dots, \text{Sen}_{N_{\text{msg}}}$, send a message while we observe the system. Using their sending profiles a corresponding sequence of receivers $\text{Rec}_1, \dots, \text{Rec}_{N_{\text{msg}}}$ is selected to receive their messages. The probability of any receiver sequence is easy to compute. We denote this matching between senders and receivers as $\mathcal{M}$:

$$\Pr[\mathcal{M}|\Psi] = \prod_{x \in [1, N_{\text{msg}}]} \Pr[\text{Sen}_x \to \text{Rec}_x | \Psi_x].$$

In parallel with the matching process where users choose their communication partners, an anonymity system $\mathcal{A}$ is used. This anonymity system is abstracted as a bipartite graph linking input messages i_x with potential output messages o_y, regardless of the identity of their senders and receivers. We note that completeness of the bipartite graph is not required by the model. The edges of the bipartite graph are weighted with w_{xy} that is simply the probability of the input message i_x being output as o_y: $w_{xy} = \Pr[i_x \to o_y | \mathcal{A}]$.

This anonymity system $\mathcal{A}$ is used to determine a particular assignment of messages according to the weights w_{xy}. A single perfect matching on the bipartite

graph described by $\mathcal{A}$ is selected to be the correspondence between inputs and outputs of the anonymity system for a particular run of the anonymity protocol. We call it the assignment of inputs to outputs and denote it by Φ. Contrary to previous work [20] on probabilistic modelling, and following the tendency started by Troncoso *et al.* [21], we consider all inputs simultaneously. In this case the probability of the assignment Φ is easy to calculate, given the set of all individual assignments $(i_x \to o_x)$:

$$\Pr[\Phi|\mathcal{A}] = \prod_x \frac{\Pr[i_x \to o_x|\mathcal{A}]}{\sum_{\text{free } i_y} \Pr[i_y \to o_x|\mathcal{A}]}.$$

This is simply the probability of the matching given the anonymity system weights. By free i_y we denote the set of sent messages i that has not yet been assigned an output message o as part of the match.

The assignment Φ of the anonymity system and the matching $\mathcal{M}$ of senders and receivers are composed to make up the observation of the adversary, that we denote as $\mathcal{O}$. An adversary observes messages from particular senders Sen_x entering the anonymity as messages i_x, and on the other side messages o_y exiting the network on their way to receivers Rec_y. No stochastic process takes place in this deterministic composition and therefore $\Pr[\mathcal{O}|\mathcal{M}, \Phi, \Psi, \mathcal{A}] = 1$.

Now that we have defined a full generative model for all the quantities of interest in the system, we turn our attention to the inference problem: the adversary observes $\mathcal{O}$ and knows about the anonymity system $\mathcal{A}$, but is ignorant about the profiles Ψ, the matching $\mathcal{M}$ and the assignment Φ. We use Bayes theorem to calculate the probability $\Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}]$. We start with the joint distribution and solve for it:

$$\begin{aligned}
\Pr[\mathcal{O}, \mathcal{M}, \Phi, \Psi|\mathcal{A}] &= \Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}] \cdot \Pr[\mathcal{O}|\mathcal{A}] \\
\Pr[\mathcal{O}, \mathcal{M}, \Phi, \Psi|\mathcal{A}] &= \Pr[\mathcal{O}|\mathcal{M}, \Phi, \Psi, \mathcal{A}] && (\equiv 1) \\
&\quad \cdot \Pr[\mathcal{M}|\Phi, \Psi, \mathcal{A}] && (\equiv \Pr[\mathcal{M}|\Psi]) \\
&\quad \cdot \Pr[\Phi|\Psi, \mathcal{A}] && (\equiv \Pr[\Phi|\mathcal{A}]) \\
&\quad \cdot \Pr[\Psi|\mathcal{A}] \\
\Rightarrow \Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}] &= \frac{\Pr[\mathcal{M}|\Psi] \Pr[\Phi|\mathcal{A}]}{\Pr[\mathcal{O}|\mathcal{A}] \equiv \mathcal{Z}} \Pr[\Psi|\mathcal{A}]
\end{aligned}$$

We have discussed how to calculate the probabilities $\Pr[\mathcal{M}|\Psi]$ and $\Pr[\Phi|\mathcal{A}]$. The quantity $\Pr[\Psi|\mathcal{A}] \equiv \Pr[\Psi]$ is the a-prior belief the attacker has about user profiles and it is independent from the chosen anonymity system $\mathcal{A}$. We consider throughout our analysis that all profiles are a-priori equally probable and reduce it to a constant $\Pr[\Psi] = c$. Taking into account those observations we conclude that the posterior probability sought is,

$$\Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}] \sim \prod_{x \in [1, N_{\text{msg}}]} \Pr[\text{Sen}_x \to \text{Rec}_x|\Psi_x] \cdot \prod_x \frac{\Pr[i_x \to o_x|\mathcal{A}]}{\sum_{\text{free } i_y} \Pr[i_y \to o_x|\mathcal{A}]}$$

where we omit the constant normalising factor $\Pr[\mathcal{O}|\mathcal{A}]$ as it is very hard to calculate, which restricts the methods we can use to manipulate the a-posterior distribution.

It is computationally unfeasible to exhaustively enumerating the states of this distribution. Hence to calculate the marginals of interest such as profiles of users, or likely recipients of specific messages, we have to resort to sampling states from that distribution. Sampling directly is very hard (due to the interrelation between the profiles, the matches and the assignments) hence Markov Chain Monte Carlo methods are used.

3.1 A Gibbs sampler for the Vida Black-box model

Sampling states $(\mathcal{M}_j, \Phi_j, \Psi_j) \sim \Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}]$ directly is hard, due to the complex interactions between the random variables. A Gibbs sampler significantly simplifies this process by only requiring us to sample from the marginal distributions of the random variables sought. Given an arbitrary initial state (Φ_0, Ψ_0) we can perform ι iterations of the Gibbs algorithm as follows:

$$\begin{aligned} &\text{for } j := 1 \dots \iota : \\ \Phi_j, \mathcal{M}_j &\sim \Pr[\Phi, \mathcal{M}|\Psi_{j-1}, \mathcal{O}, \mathcal{A}] \\ \Psi_j &\sim \Pr[\Psi|\Phi_j, \mathcal{M}_j, \mathcal{O}, \mathcal{A}] \quad . \end{aligned}$$

Each of these marginal probabilities distributions is easy to sample:

– The distribution of assignments $\Pr[\Phi, \mathcal{M}|\Psi_{j-1}, \mathcal{O}, \mathcal{A}]$ is subtle to sample directly. Each message assignment $i_x \to o_x$ has to be sampled, taking into account that some message assignments are already taken by the time input message i_x is considered. For each input message i_x we sample an assignment o_y according to the distribution:

$$\begin{aligned} i_x \to o_y &\sim \Pr[i_x \to o_y | \text{free } o_y, \forall_{\text{assigned } o_v} i_v \to o_v, \mathcal{A}, \Psi] \\ &= \frac{\Pr[i_x \to o_y|\mathcal{A}] \cdot \Pr[\text{Sen}_x \to \text{Rec}_y|\Psi_x]}{\sum_{\text{free } o_y} \Pr[i_x \to o_y|\mathcal{A}] \cdot \Pr[\text{Sen}_x \to \text{Rec}_y|\Psi_x]} \quad . \end{aligned}$$

For complex anonymity systems $\mathcal{A}$, this algorithm might return only partial matches, when at some point an input message i_x has no unassigned candidate output message o_y left. Since we are only interested in perfect matchings, where all input messages are matched with different output messages, we reject such partial states and re-start the sampling of the assignment until a valid perfect matching is returned. This is effectively a variant of rejection sampling, to sample valid assignments.
The matchings between senders and receivers are uniquely determined by the assignments and the observations, so we can update them directly without any need for sampling, and regardless of the profiles (i.e. $\mathcal{M}_j = f(\Psi_j, \mathcal{O})$).

– The distribution of profiles $\Pr[\Psi|\Phi_j, \mathcal{M}_j, \mathcal{O}, \mathcal{A}]$ is straightforward to sample given the matching $\mathcal{M}_j$ and assuming that individual profiles Ψ_x are multinomial distributions.
We note that the Dirichlet distribution is a conjugate prior of the multinomial distribution, and we use it to sample profiles for each user. We denote as $\Psi_x = (\Pr[\text{Sen}_x \to \text{Rec}_1], \ldots, \Pr[\text{Sen}_x \to \text{Rec}_{N_{\text{user}}}])$ the multinomial profile of user Sen_x. We also define a function that counts the number of times a user Sen_x is observed sending a message to user Rec_y in the match $\mathcal{M}$, and denote it as $\text{Ct}_{\mathcal{M}}(\text{Sen}_x \to \text{Rec}_y)$. Sampling profiles $(\Psi_1, \ldots, \Psi_{N_{\text{user}}}) \sim \Pr[\Psi|\mathcal{M}]$ involves sampling independently each sender's profile Ψ_x separately from a Dirichlet distribution with the following parameters:

$$\Psi_x \sim \text{Dirichlet}(\text{Ct}_{\mathcal{M}}(\text{Sen}_x \to \text{Rec}_1) + 1, \ldots, \text{Ct}_{\mathcal{M}}(\text{Sen}_x \to \text{Rec}_{N_{\text{user}}}) + 1) \quad .$$

If the anonymity system $\mathcal{A}$ describes a simple bipartite graph, the rejection sampling algorithm described can be applied to sample assignments $i_x \to o_x$ for all messages. When this variant of rejection sampling becomes expensive, due to a large number of rejections, a Metropolis-Hastings [4] based algorithm can be used to sample perfect matchings on the bipartite graph according to the distribution $\Pr[\Phi, \mathcal{M}|\Psi_{j-1}, \mathcal{O}, \mathcal{A}]$. Our implementation was tested against mix-based anonymity systems, with bipartite graphs representing the anonymity system that do not lead to any rejections.

The Gibbs sampler can be run multiple times to extract multiple samples from the a-posterior distribution $\Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}]$. Instead of restarting the algorithm at an arbitrary state $(\mathcal{M}_0, \Phi_0, \Psi_0)$, it is best to set the starting state to the last extracted sample, that is likely to be within the typical set of the distribution. This speeds up convergence to the target distribution.

4 A computationally simple Vida Red-Blue model

After the PMDA [21] it has become dogma that sender profiles have to be co-estimated simultaneously with the assignments, and our Bayesian analysis so far reflects this approach. Senders are associated with multinomial profiles with which they choose specific correspondents. We sample these profiles using the Dirichlet distribution, and use them to directly sample weighted perfect assignments in the anonymity system. The output of the algorithm is a set of samples of the hidden state, that allows the adversary to estimate the marginal distributions of specific senders sending to specific receivers.

We note that this approach is very generic, and might go beyond the day to day needs of a real-world adversary. An adversary is likely to be interested in particular target senders or receivers, and might want to answer the question: "who has sent this message to Bob?" or "who is friends with receiver Bob?". We present the Vida Red-Blue model to answer such questions, which is much simpler, both mathematically and computationally, than the generic Vida model presented so far.

Consider that the adversary chooses a target receiver Bob (that we call "Red"), while ignoring the exact identity of all other receivers and simply tagging them as "Blue". The profiles Ψ_x of each sender can be collapsed into a simple binomial distribution describing the probability sender x sends to Red or to Blue. It holds that:

$$\Pr[\text{Sen}_x \to \text{Red}|\Psi_x] + \Pr[\text{Sen}_x \to \text{Blue}|\Psi_x] = 1. \tag{1}$$

Matchings $\mathcal{M}$ map each observed sender of a message to a receiver class, either Red or Blue. Given the profiles Ψ the probability of a particular match $\mathcal{M}$ is:

$$\Pr[\mathcal{M}|\Psi] = \prod \Pr[\text{Sen}_x \to \text{Red / Blue}|\Psi_x]$$

The real advantage of the Vida Red-Blue model is that different assignments Φ now belong to equivalence classes, since all Red or Blue receivers are considered indistinguishable from each other. In this model the assignment bipartite graph can be divided into two sub-graphs: the sub-graph Φ_R contains all edges ending on the Red receiver (as she can receive more than one message in a mixing round), while the sub-graph Φ_B contains all edges ending on a Blue receiver. We note that these sub-graphs are complementary and any of them uniquely defines the other. The probability of each Φ can then be calculated as:

$$\begin{aligned}
\Pr[\Phi|\mathcal{A}] &= \sum_{\forall \Phi_B} \Pr[\Phi_B, \Phi_R|\mathcal{A}] = \\
&= \sum_{\forall \Phi_B} \Pr[\Phi_B|\Phi_R, \mathcal{A}] \cdot \Pr[\Phi_R|\mathcal{A}] = \\
&= \Pr[\Phi_R|\mathcal{A}] \cdot \sum_{\forall \Phi_B} \Pr[\Phi_B|\Phi_R, \mathcal{A}] = \\
&= \Pr[\Phi_R|\mathcal{A}]
\end{aligned}$$

The probability of an assignment in an equivalence class defined by the assignment to Red receivers, only depends on Φ_R describing this assignment. The probability of assignment Φ_R can be calculated analytically as:

$$\Pr[\Phi_R|\mathcal{A}] = \prod_{x \in \Phi_R} \frac{\Pr[i_x \to o_x]}{\sum_{\text{free } i_j} \Pr[i_j \to o_x]}.$$

The assignment Φ_R must be a sub-graph of at least one perfect matching on the anonymity system $\mathcal{A}$, otherwise the probability becomes $\Pr[\Phi|\mathcal{A}] = 0$. As for the full model the probability of all the hidden quantities given the observation is:

$$\Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}] = \frac{\Pr[\mathcal{M}|\Psi] \Pr[\Phi_R|\mathcal{A}]}{\Pr[\mathcal{O}|\mathcal{A}] \equiv \mathcal{Z}} \Pr[\Psi|\mathcal{A}] \tag{2}$$

The a-prior probability over profiles $\Pr[\Psi|\mathcal{A}]$ is simply a prior probability over parameters of a binomial distribution. Each profile can be distributed as $\Pr[\Psi_x|\mathcal{A}] =$

$\text{Beta}(1,1)$ if nothing is to be assumed about the sender's x relationship with the Red receiver.

In practice a prior distribution $\Pr[\Psi_x|\mathcal{A}] = \text{Beta}(1,1)$ is too general, and best results are achieved by using a prior supporting skewed distributions, such as $\text{Beta}(1/100, 1/100)$. This reflects the fact that social ties are a-prior either strong or non existent. Given enough evidence the impact of this choice of prior fades quickly away.

4.1 A Gibbs sampler for the Vida Red-Blue model

Implementing a Gibbs sampler for the Vida Red-Blue model is very simple. The objective of the algorithms is, as for the general model, to produce samples of profiles (Ψ_j), assignments and matches ($\Phi_j, \mathcal{M}_j$) distributed according to the Bayesian a-posterior distribution $\Pr[\mathcal{M}, \Phi, \Psi|\mathcal{O}, \mathcal{A}]$ described by eq. 2.

The Gibbs algorithm starts from an arbitrary state (Ψ_0, Φ_0) and iteratively samples new marginal values for the profiles $(\Phi_j, \mathcal{M}_j \sim \Pr[\Phi, \mathcal{M}|\Psi_{j-1}, \mathcal{O}, \mathcal{A}])$ and the valid assignments $(\Psi_j \sim \Pr[\Psi|\mathcal{M}_j, \Phi_j, \mathcal{O}, \mathcal{A}])$. The full matchings are a deterministic function of the assignments and the observations, so we can update them directly without any need for sampling (i.e. $\mathcal{M}_j = f(\Psi_j, \mathcal{O})$).

As for the general Gibbs sampler, sampling from the desired marginal distributions can be done directly. Furthermore the Vida Red-Blue model introduces some simplifications that speed up inference:

- **Sampling assignments.** Sampling assignments of senders to Red nodes (i.e. $\Phi_{Rj}, \mathcal{M}_j \sim \Pr[\Phi, \mathcal{M}|\Psi_{j-1}, \mathcal{O}, \mathcal{A}]$) can be performed by adapting the rejection sampling algorithm presented for the general model. The key modification is that only assignments to Red receivers are of interest, and only an arbitrary assignment to blue receivers is required (to ensure such an assignment exists). This time for each Red output messages o_x we sample an input message i_x according to the distribution:

$$\begin{aligned} i_x \rightarrow o_y \sim & \Pr[i_x \rightarrow o_y | \text{free } i_x, \forall_{\text{assigned } i_v} i_v \rightarrow o_v, \mathcal{A}, \Psi] \\ = & \frac{\Pr[i_x \rightarrow o_y|\mathcal{A}] \cdot \Pr[\text{Sen}_x \rightarrow \text{Red}|\Psi_x]}{\sum_{\text{free } i_j} \Pr[i_j \rightarrow o_y|\mathcal{A}] \cdot \Pr[\text{Sen}_j \rightarrow \text{Red}|\Psi_x]} \end{aligned}$$

- **Sampling profiles.** Sampling a profile $\Psi_j \sim \Pr[\Psi|\mathcal{M}_j, \Phi_j, \mathcal{O}, \mathcal{A}]$ for every user x simply involves drawing a sample from a Beta distribution with parameters related to the number of links to Blue and Red receivers. To be formal we define a function $\text{Ct}_{\mathcal{M}}(\text{Sen}_x \rightarrow \text{Red, Blue})$ that counts the number of messages in a match that a user x sends to a Red or Blue receiver. The profile of user x is then sampled as:

$$\Psi_x \sim \text{Beta}(\text{Ct}_{\mathcal{M}}(\text{Sen}_x \rightarrow \text{Blue}) + 1, \text{Ct}_{\mathcal{M}}(\text{Sen}_x \rightarrow \text{Red}) + 1)$$

  This yields a binomial parameter that is the profile of user x, describing the probability they send a message to a Red target user.

The cost of each iteration is proportional to sampling N_{user} Beta distributions, and sample from the distribution of senders of each of the Red messages. Both the sampling of profiles, and the sampling of assignments can be performed in parallel, depending on the topology. In case a large number of samples are needed multiple Gibbs samplers can be run on different cores or different computers to produce them.

4.2 Evaluation

The Vida Red-Blue model for inferring user profiles and assignments was evaluated against synthetic anonymized communication traces, to test its effectiveness. The communication traces include messages sent by up to 1000 senders to up to 1000 receivers. Each sender is assigned 5 contacts at random, to whom they send messages with equal probability. Messages are anonymized in discrete rounds using a threshold mix that gathers 100 messages before sending them to their receivers as a batch.

The generation of communication patterns was peculiar to ensure a balance between inferring the communications of a target user (as in the traditional disclosure, hitting set and statistical disclosure attacks) to a designated Red receiver, as well as to gain enough information about other users to build helpful profiles for them. A target sender was included in 20% of the rounds, and the Red node was chosen to be one of their friends. A sequence of experiments were performed to assess the accuracy of the attack after observing an increasing number of rounds of communication.

The aim of each experiment is to use the samples returned by a Gibbs sampler implementing the Vida Red-Blue model to guess the sender of each message that arrives at a designated Red receiver. The optimal Bayes criterion [2] is used to select the candidate sender of each Red message: the sender with the highest a-posterior probability is chosen as the best candidate. This probability is estimated by counting the number of times each user were the sender of a target Red message in the samples returned by the Gibbs algorithm. The Bayesian probability of error, i.e. the probability another sender is responsible for the Red message, is also extracted, as a measure of the certainty of each of these "best guesses". For each experiment the Gibbs sampler was used to extract 200 samples, using 100 iterations of the Gibbs algorithm each. The first 5 samples were discarded, to ensure stability is reached before drawing any inferences.

A summary of the results for each experiment is presented in Figure 2. The top graph illustrates the fraction of correct guesses per experiment (on the x axis – we selected 20 random experiments to display per round number) grouped by the number of rounds of communication observed (16, 32, 64, 128, 256, 512 and 1024). For each experiment the fraction of correctly identified senders is marked by a circle, along with its 90% confidence interval. The dashed line of the same graph represents the prediction of success we get from the Bayesian probability of error. The bottom graph on Figure 2 illustrates on a logarithmic scale the inferred probability assigned to the Red node for the target sender, for each of the experiments. The experiments for which a high value of this probability

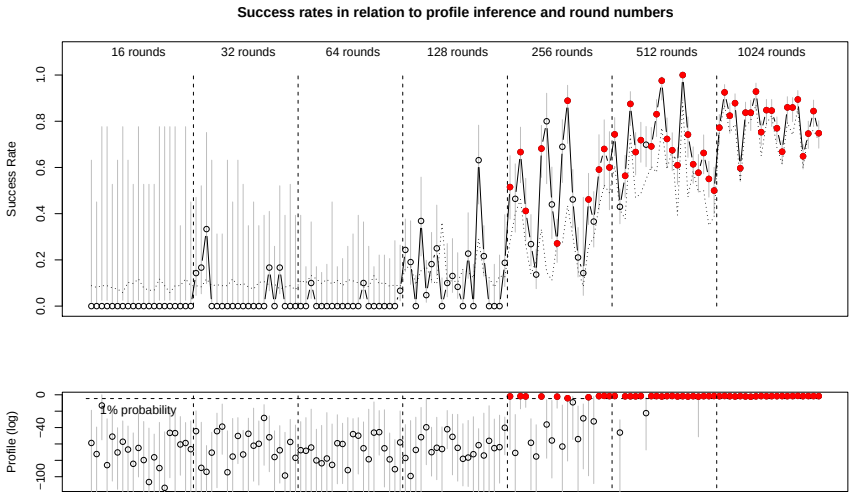

Fig. 2. Performance of the Vida Red-Blue model in assigning senders to the target red receiver, as a function of the number of rounds observed. Twenty sample experiments are used per round number.

are inferred (median greater than 1%) are marked by a solid red circle on both graphs. The 50% confidence interval over the profile parameter is also plotted.

Some key conclusions emerge from the experiments illustrated on Figure 2:

- The key trend we observe is, as expected, that the longer the observation in terms of rounds, the better the attack. Within 1024 rounds we expect the target sender to have sent about 40 messages to the designated red target. Yet, the communication is traced to them on average 80% of the cases with high certainty. Even when only 256 rounds are observed the correct assignment is guessed in about 50% of the time.
- The quality of the inference when it comes to the correspondence between messages, senders and receivers, is intimately linked to the quality of the profile inference. The solid red circles mark experiments that concluded that the median value for the probability the target sender is friends with the target Red receiver is high (greater than 1%). We observe that these experiments are linked to high success rates when it comes to linking individual messages to the target sender. We also observe the converse: insufficient data leads to poor profiles, that in turn lead to poor predictions about communication relationships.
- The probability of success estimates (represented on the top graph by a dotted line) predict well the success rate of the experiments. Our prediction systematically falls within the 90% confidence interval of the estimated error

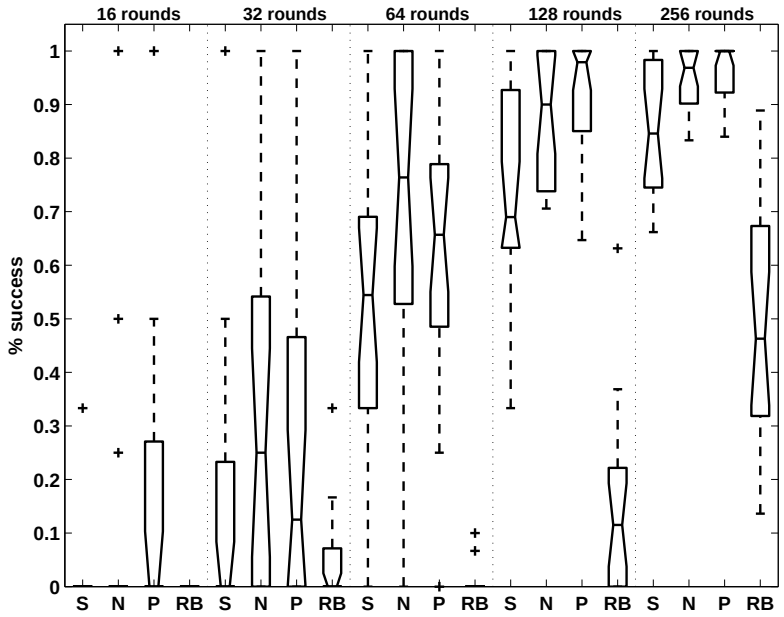

Fig. 3. Performance of the Vida Red-Blue inference model (RB) compared to the SDA (S), NSDA (N) and PMDA (P.)

rate. This shows that the Vida Red-Blue model is a good representation of the process that generated the traces and thus the estimates coincide with the actual observed error rate, on average. This is due to the very generic model for Vida Red-Blue profiles that represent reality accurately after a few rounds. Yet, when few rounds are observed the a-prior distribution of profiles dominates the inference, and affects the error estimates.

A key question is how the results from the Vida Red-Blue model compare with traditional traffic analysis attacks, like the SDA [15], the NSDA [21] or the PMDA [21]. The SDA attack simply uses first order frequencies to guess the profiles of senders. It is fast but inaccurate. The normalised SDA (NSDA) constructs a traffic matrix from senders to receivers, that is normalised to be doubly stochastic. The operation is as fast as matrix multiplication, and yields very good results. The PMDA finds perfect matchings between senders and receivers based on a rough profile extraction step – it is quite accurate but slow.

Figure 3 illustrates the relative performances of the different attacks compared with the Vida Red-Blue model proposed. We observe that the inference based technique is quite competitive, against the SDA, but performs worse than the NSDA and PMDA in most settings. This is due to our strategy for extracting best estimates for the senders: we use the output samples to chose the sender with highest marginal probability instead of extracting a full match with the

maximal marginal probability. In that sense applying an algorithm to find the maximal perfect matching based on the marginal probabilities output by the RB attacks should produce much better results.

Despite the lower success rate inference based techniques can be advantageous. Their key strength is the certainty that no systematic bias has been introduced by reusing data twice, as reported in [10, 21], and the tangible and reliable error estimate they output. A traffic analyst is thus able to judge the quality of the inference to guide them operationally.

A second important advantage is the ability to infer who is the "second most likely" receiver, compute anonymity metrics, or other arbitrary statements on the a-posterior probability distribution of profiles and assignments. This can be done efficiently simply using the samples output by the Gibbs algorithm. Furthermore the correct probabilities of error can be associated with those probabilistic statements.

5 Discussion & future directions

The Bayesian treatment of long term attacks against anonymity systems is promising, but still at its infancy. We foresee some key theoretical as well as implementation steps to move the state of the art forward.

- **Bipartite weighted anonymity set**. The Vida Black-box model as well as the Vida Red-Blue model proposed represent an observation from an anonymity system as a generic weighted bipartite graph, linking senders with receivers. Our experiments, on the other hand, only considered anonymity systems working in discrete rounds, forming full bipartite sub-graphs with a number of senders equal to the batch size. This is a limitation of our sampler implementation, that could be extended to deal with the general case of any bipartite weighted network.
  While in theory this modification is straightforward, in practice it is harder to sample directly matchings from arbitrary bipartite graphs. The rejection sampling algorithm suggested can be inefficient, since it might use links that are not part of a perfect matching, forcing multiple aborts. It might be wise to first prune the assignment graph from such edges using techniques from the constrain satisfaction literature such as Regin's algorithm [18].
- **Profile models**. The a-prior model for user profiles is very generic, meaning that it can represent, and thus learn, any multinomial distribution of receivers per sender. While being generic more information could be incorporated if it is established that the profile belongs to a social network (with some standard characteristics like degree, clustering etc). Traditional hitting set as well as disclosure attacks make extensive use of the number of friends of a target sender to be applicable at all, whereas the presented approaches do not require such information. Yet, adding related constraints would yield better results.
- **Learning social networks**. It has been an open problem in the literature how to incorporate known information about communication patterns

to help the inference of unknown communication patterns, and some ad-hoc techniques were presented to combine social network information to de-anonymize traces, along with a discussion of systematic errors introduced [10]. The sampling techniques presented in this work can be straightforwardly modified to incorporate known correspondences between senders and receivers: the Gibbs sampler is modified to only sample valid assignments that contain the known matches. These known assignments, far from being useless, drive the sampling of profiles (as part of the Gibbs sampling) leading to higher quality profiles, which in turn become higher quality assignments for the unknown messages.

– **Beyond communications.** Both models presented are very generic and apply to attempts to anonymize traces that are not communications. As long as a system has users with multinomial preferences, that are expressed and anonymized in an arbitrary manner (as long as there is one expressed preference per observed action), our algorithms are applicable to de-anonymize the preferences and extract user profiles. This problem has recently received considerable attention though de-anonymization algorithms applied to the NetFlix database [16].

6 Conclusions

The contribution of this work is two-fold: First it presents Vida, the first truly general model for abstracting any anonymity system, in the long term, to perform de-anonymization attacks. Users and their preferences are modelled in the most generic way, using multinomial profile, eliminating the need to know the number of contacts each sender has. Instead of abstracting an anonymity system as a single threshold mix, or even pool mix, an arbitrary weighted mapping of input to output messages can be used. We show that the model performs well when it comes to guessing who is talking to whom, as well as guessing the profiles of senders. The Vida Red-Blue model focuses on the need the working traffic analyst has to infer patterns of communications to specific targets – it has the potential to be implemented efficiently and parallelized aggressively.

The second contribution is methodological, and might be even more significant than the specific Vida models. We demonstrate that probabilistic modelling, Bayesian inference, and the associated conceptual toolkit relating to Monte Carlo Markov chain sampling is an appropriate framework upon which to build traffic analysis attacks. It ensures that information is used properly avoiding over fitting or systematic biases; it provides a clear framework to perform the analysis starting with the definition of a probabilistic model, that is inverted and sampled to estimate quantities of interest; it provides good and clear estimates of error, as well as the ability to answer arbitrary questions about the hidden state with a clear probability statement. These qualities are in sharp contrast with the state of the art in traffic analysis, that provides ad-hoc best guesses of very specific quantities, with a separate analysis to establish their accuracy based on labeled data – something that the traffic analyst does not have on the ground.

We hope this work is the start of an exploration of the applicability of inference techniques to problems in traffic analysis – that will eventually outperform established techniques. Some clear future directions include the definition of better user models, the analysis of the internals of anonymity systems, as well as a better integration of prior information and learning. The inference approach leans itself well to be extended to encompass these problems, that have in the past been a thorn on the side of traffic analysis techniques.

Acknowledgements. The authors would like to thank the participants of the second UK anonymity meet-up in 2008, and in particular Andrei Serjantov, Ben Laurie, and Tom Chothia for their valuable comments on this research. While this work was developed its direction benefited considerably by our discussions with Steven Murdoch, who also provided significant logistical support. C. Troncoso is a research assistant of the Fund for Scientific Research in Flanders (FWO) and this work was partly performed while C. Troncoso was an intern at Microsoft Research Cambridge, between September and December 2008. This work was supported in part by the IAP Programme P6/26 BCRYPT of the Belgian State.

References

1. Dakshi Agrawal, Dogan Kesdogan, and Stefan Penz. Probabilistic Treatment of MIXes to Hamper Traffic Analysis. In *Proceedings of the 2003 IEEE Symposium on Security and Privacy*, pages 16–27, May 2003.
2. Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Probability of error in information-hiding protocols. In *Proceedings of the 20th IEEE Computer Security Foundations Symposium (CSF20)*, 2007.
3. David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 4(2), February 1981.
4. Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.
5. George Danezis. Statistical disclosure attacks: Traffic confirmation in open environments. In Gritzalis, Vimercati, Samarati, and Katsikas, editors, *Proceedings of Security and Privacy in the Age of Uncertainty, (SEC2003)*, pages 421–426, Athens, May 2003. IFIP TC11, Kluwer.
6. George Danezis and Claudia Diaz. A survey of anonymous communication channels. Technical Report MSR-TR-2008-35, Microsoft Research, January 2008.
7. George Danezis, Claudia Diaz, and Carmela Troncoso. Two-sided statistical disclosure attack. In Nikita Borisov and Philippe Golle, editors, *Proceedings of the Seventh Workshop on Privacy Enhancing Technologies (PET 2007)*, Ottawa, Canada, June 2007. Springer.
8. George Danezis and Andrei Serjantov. Statistical disclosure or intersection attacks on anonymity systems. In *Proceedings of 6th Information Hiding Workshop (IH 2004)*, LNCS, Toronto, May 2004.
9. Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In Roger Dingledine and Paul Syverson, editors, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*. Springer-Verlag, LNCS 2482, April 2002.

10. Claudia Diaz, Carmela Troncoso, and Andrei Serjantov. On the impact of social network profiling on anonymity. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth International Symposium on Privacy Enhancing Technologies (PETS 2008)*, pages 44–62, Leuven, Belgium, July 2008. Springer.
11. Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, July 2003.
12. Dogan Kesdogan, Dakshi Agrawal, and Stefan Penz. Limits of anonymity in open environments. In Fabien Petitcolas, editor, *Proceedings of Information Hiding Workshop (IH 2002)*. Springer-Verlag, LNCS 2578, October 2002.
13. Dogan Kesdogan and Lexi Pimenidis. The hitting set attack on anonymity protocols. In *Proceedings of 6th Information Hiding Workshop (IH 2004)*, LNCS, Toronto, May 2004.
14. David J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
15. Nick Mathewson and Roger Dingledine. Practical traffic analysis: Extending and resisting statistical disclosure. In *Proceedings of Privacy Enhancing Technologies workshop (PET 2004)*, volume 3424 of *LNCS*, pages 17–34, May 2004.
16. Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125. IEEE Computer Society, 2008.
17. Jean-François Raymond. Traffic Analysis: Protocols, Attacks, Design Issues, and Open Problems. In H. Federrath, editor, *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*, pages 10–29. Springer-Verlag, LNCS 2009, July 2000.
18. Jean-Charles Régin. A filtering algorithm for constraints of difference in csps. In *AAAI*, pages 362–367, 1994.
19. Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Roger Dingledine and Paul Syverson, editors, *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)*. Springer-Verlag, LNCS 2482, April 2002.
20. Vitaly Shmatikov. Probabilistic analysis of an anonymity system. *Journal of Computer Security*, 12(3-4):355–377, 2004.
21. Carmela Troncoso, Benedikt Gierlichs, Bart Preneel, and Ingrid Verbauwhede. Perfect matching disclosure attacks. In Nikita Borisov and Ian Goldberg, editors, *Proceedings of the Eighth International Symposium on Privacy Enhancing Technologies (PETS 2008)*, volume 5134 of *Lecture Notes in Computer Science*, pages 2–23, Leuven,BE, 2008. Springer-Verlag.